Descriptive Statistics using RStudio

Edre MA, DrPH

2020-12-16

```
# ======================
# Descriptive Statistics
# R Biostat Workshop IIUM
# Edre MA, DrPH
# ======================



#You you are a researcher involved in a hypertension study
#objective 1: To describe the background characteristics of respondents
#objective 2: To determine the prevalence of hypertension
#objective 3: To determine the factors contributing to hypertension

#libraries needed to be installed

#readr
#smartEDA
#moments
#ggpubr
#usingR
#car
#ggplot2
#dplyr

# data

#pulling the data from GitHub

#go to https://github.com/adilzainal/IIUM_Biostatistic_workshop
#click "code" -> "Download ZIP"
#extract the ZIP file using WinRAR
#Create a new specific folder to store all files in your desktop
#set as working directory

#loading the data

#if csv (.csv)
library(readr)

## Warning: package 'readr' was built under R version 3.6.3

healthstat <- read_csv("healthstatus6.csv") #load the file and make as object
```

```
## Parsed with column specification:
## cols(
##   id = col_double(),
##   age = col_double(),
##   sex = col_character(),
##   exercise = col_character(),
##   smoking = col_character(),
##   wt = col_double(),
##   ht = col_double(),
##   sbp = col_double(),
##   dbp = col_double(),
##   hba1c = col_double(),
##   hcy = col_double(),
##   wt2 = col_double(),
##   wt3 = col_double(),
##   sbp2 = col_double(),
##   sbp3 = col_double(),
##   dbp2 = col_double(),
##   dbp3 = col_double()
## )
```

```
View(healthstat)

#objective 1: To describe the background characteristics of respondents
#summarising numerical values

# we choose 3 IVs: age,sbp,dbp

library(SmartEDA)
```

```
## Warning: package 'SmartEDA' was built under R version 3.6.3

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
ExpCustomStat(healthstat,
              Nvar=c("age","sbp","dbp"),
              stat = c('mean', 'sd', 'median', 'IQR'))
```

```
##    Attribute      mean       sd median IQR
## 1:       age  42.16340 8.932096     42  11
## 2:       sbp 132.24837 7.956527    132  13
## 3:       dbp  86.53595 6.268159     87   9
```

```
#normality assumption check

#there are 5 criteria before you make decision what to report:

#1.mean~median

ExpCustomStat(healthstat,
              Nvar=c("age","sbp","dbp"),
              stat = c('mean','median'))
```

```
##    Attribute      mean median
## 1:       age  42.16340     42
## 2:       sbp 132.24837    132
## 3:       dbp  86.53595     87
```

```
#2. acceptable skewness & kurtosis +-2d

library(moments)
ExpCustomStat(healthstat,
              Nvar=c("age","sbp","dbp"),
              stat = c('skewness','kurtosis'))
```

```
##    Attribute    skewness kurtosis
## 1:       age  0.16179220 2.783220
## 2:       sbp  0.22172135 2.417301
## 3:       dbp -0.02148621 2.548945
```

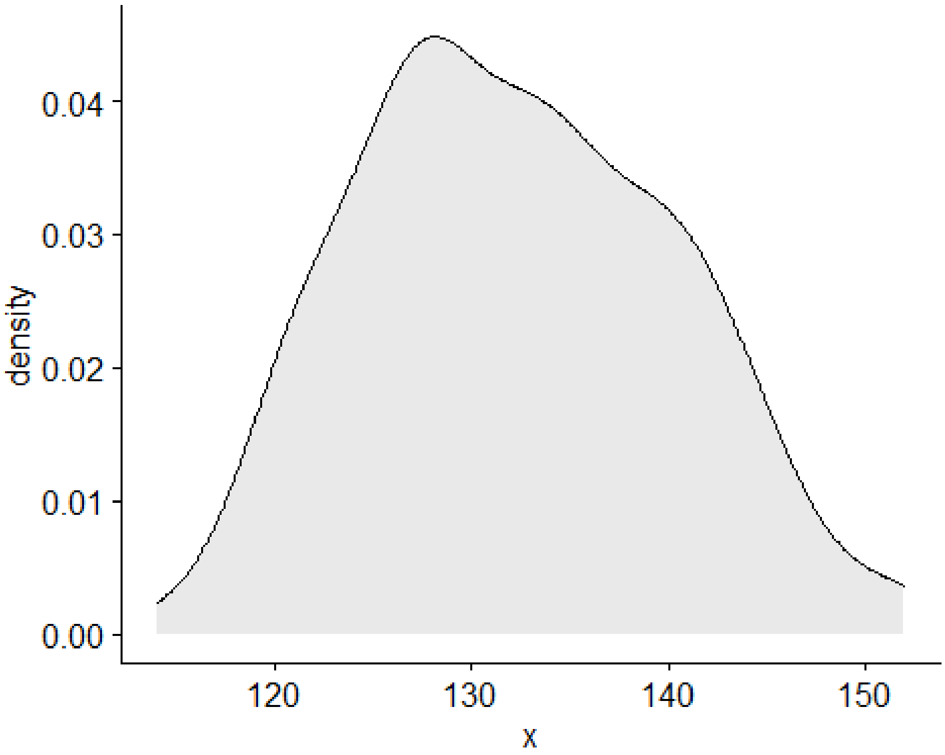
```
#3. bell shaped curve (The MOST powerful determinant of normality)
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 3.6.3
```

```
## Loading required package: ggplot2
```

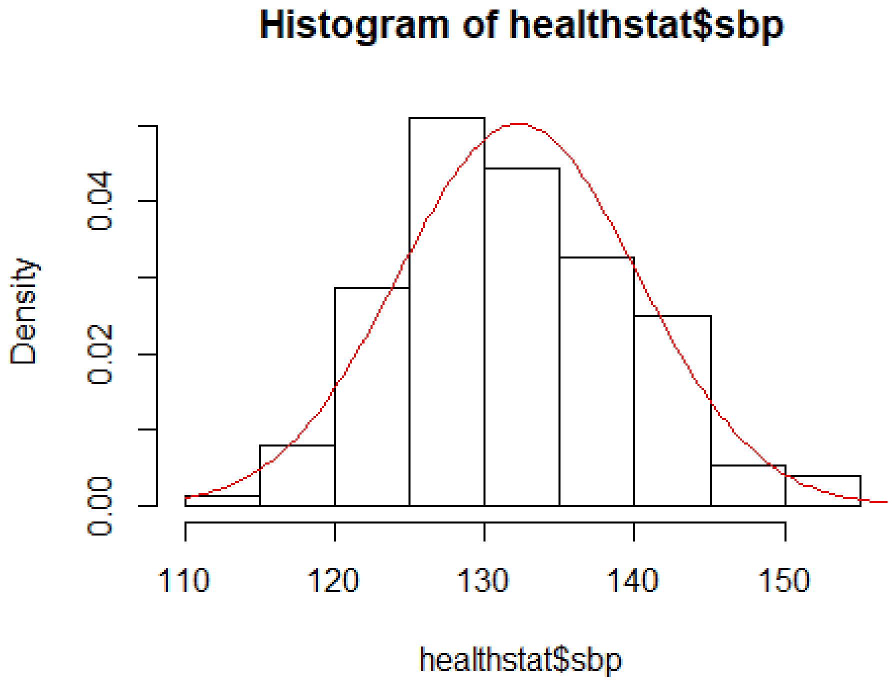
```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

```
ggdensity(healthstat$sbp, fill = "lightgray")
```

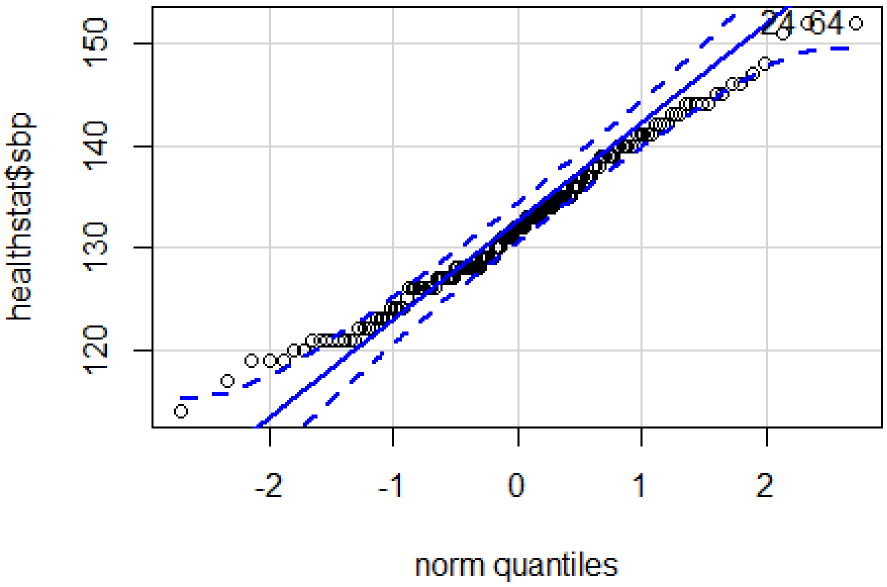


```
library(UsingR)

## Warning: package 'UsingR' was built under R version 3.6.3

## Loading required package: MASS

## Loading required package: HistData

## Warning: package 'HistData' was built under R version 3.6.3

## Loading required package: Hmisc

## Warning: package 'Hmisc' was built under R version 3.6.3

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Warning: package 'Formula' was built under R version 3.6.3

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
##
## Attaching package: 'UsingR'

## The following object is masked from 'package:survival':
##
##     cancer

hist(healthstat$sbp, freq = FALSE)
x <- seq(110, 160, length.out=170)
y <- with(healthstat, dnorm(x, mean(sbp), sd(sbp)))
lines(x, y, col = "red")
```

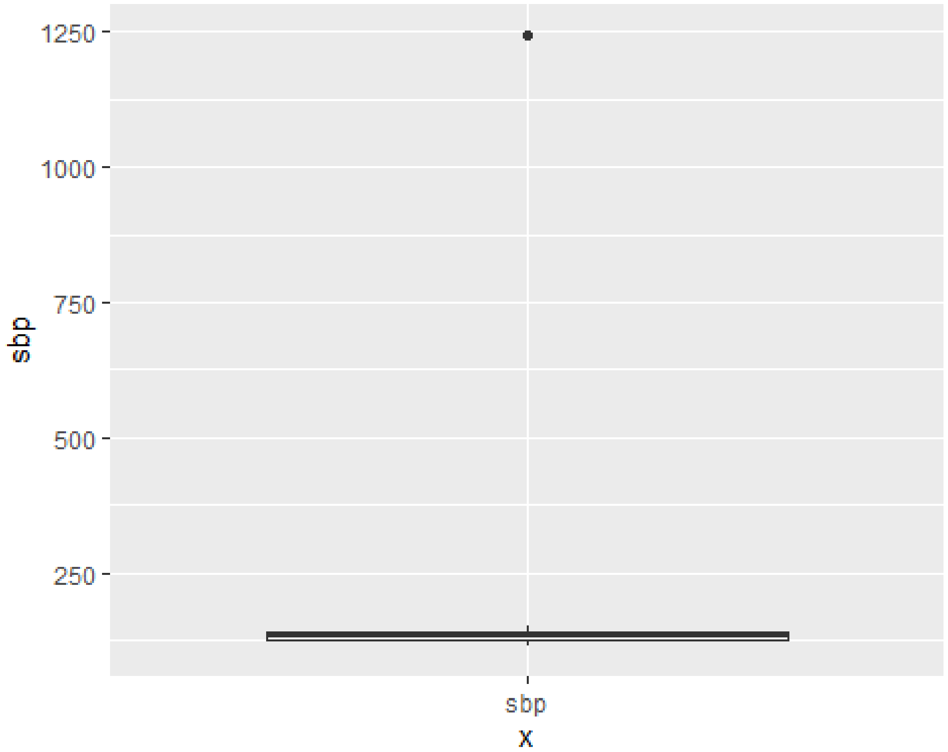


```
#4. qqplot

library(car)

## Warning: package 'car' was built under R version 3.6.3

## Loading required package: carData

qqPlot(healthstat$sbp)
```



```
## [1] 24 64

#5. normality test
shapiro.test(healthstat$sbp) #sample size less than 50

##
##  Shapiro-Wilk normality test
##
## data:  healthstat$sbp
## W = 0.98403, p-value = 0.07418

ks.test(x, "pnorm", mean=mean(healthstat$sbp), sd=sd(healthstat$sbp))

##
##  One-sample Kolmogorov-Smirnov test
##
## data:  x
## D = 0.25558, p-value = 4.523e-10
## alternative hypothesis: two-sided

#finally, make your decision
ExpCustomStat(healthstat,
              Nvar=c("age","sbp","dbp"),
              stat = c('mean', 'sd'))

##    Attribute      mean       sd
## 1:       age  42.16340 8.932096
## 2:       sbp 132.24837 7.956527
## 3:       dbp  86.53595 6.268159

#summarising categorical values

ExpCustomStat(healthstat,
              Cvar=c("sex", "smoking"),
              gpby=FALSE)

##     Level Group_by Count  Prop
## 1:   Male      sex    83 54.25
## 2: Female      sex    70 45.75
## 3:    Yes  smoking    63 41.18
## 4:     No  smoking    90 58.82

#count refers to the frequency, n
#proportion here refers to the percentage distribution of that category

#missing data

#usually coded as "NA" in the dataset
#we create a dummy object first to showcase this exercise
missing <- healthstat
missing[missing$id==57, "sbp"] <- NA

#demonstrating the row to show the missing value using dummy data
missing$sbp

##   [1] 123 122 136 127 151 128 146 145 134 122 124 138 127 145 138 126 122
128
##  [19] 135 117 147 135 139 152 126 121 132 139 137 144 135 141 130 131 144
129
##  [37] 126 127 136 123 124 121 127 131 134 124 139 128 127 132 143 128 130
144
##  [55] 124 141  NA 135 121 140 142 128 146 152 144 142 132 137 126 133 128
141
##  [73] 126 119 125 130 131 140 123 120 127 126 119 140 121 134 133 131 129
128
##  [91] 140 139 143 129 126 133 136 128 134 132 140 137 140 135 127 128 128
143
## [109] 133 119 126 132 133 131 126 140 136 135 128 141 139 135 137 132 114
121
## [127] 122 121 142 133 133 142 129 129 141 129 139 148 121 133 131 128 144
134
## [145] 123 126 120 138 135 127 124 134 121

which (is.na(missing$sbp))

## [1] 57

#outlier detection
```

```
#create an outlier dummy data
outlierdummy <- healthstat
outlierdummy[outlierdummy$id==131, "sbp"] <- 1244

#visual method
library(ggplot2)
ggplot(outlierdummy, aes(x = "sbp", y = sbp)) + geom_boxplot()
```



```
#data row method

is_outlier <- outlierdummy$sbp > 250 | outlierdummy$sbp < 70
is_outlier

##   [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FA
LSE
##  [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FA
LSE
##  [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FA
LSE
##  [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FA
LSE
##  [49] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FA
LSE
##  [61] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FA
LSE
##  [73] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FA
LSE
```

```
##  [85] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FA
LSE
##  [97] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FA
LSE
## [109] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FA
LSE
## [121] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FA
LSE
## [133] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FA
LSE
## [145] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
#objective 2: To determine the prevalence of hypertension
#objective 3: To determine the factors contributing to hypertension
#basic data transformation:categorizing

library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:car':
##
##     recode

## The following objects are masked from 'package:Hmisc':
##
##     src, summarize

## The following object is masked from 'package:MASS':
##
##     select

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#hypertension status (either sbp or dbp equal or more than 140/90mmHg, respec
tively, considered hypertensive)
#to answer objective 2

healthstatcat<-healthstat %>%
  mutate(hpt=if_else(healthstat$sbp<140 & healthstat$dbp<90,'normal','high'))

View(healthstatcat)

ExpCustomStat(healthstatcat,
              Cvar="hpt",
              stat=c("count","prop"))
```

```
##       hpt count  prop
## 1: normal    94 61.44
## 2:   high    59 38.56
```

```
#to make data preparation for objective 3
#glucose control (6.5% and above considered poor)


healthstatcat$glucontrol<-cut(healthstatcat$hba1c,
                           breaks=c(-Inf,6.49,Inf),
                           labels=c("Good", "Poor"))
summary(healthstatcat)
```

```
##        id            age             sex              exercise        
##  Min.   :  1   Min.   :21.00   Length:153         Length:153        
##  1st Qu.: 39   1st Qu.:36.00   Class :character   Class :character  
##  Median : 77   Median :42.00   Mode  :character   Mode  :character  
##  Mean   : 77   Mean   :42.16                                        
##  3rd Qu.:115   3rd Qu.:47.00                                        
##  Max.   :153   Max.   :64.00                                        
##    smoking                wt              ht             sbp       
##  Length:153         Min.   :42.60   Min.   :140.0   Min.   :114.0  
##  Class :character   1st Qu.:55.40   1st Qu.:148.0   1st Qu.:126.0  
##  Mode  :character   Median :59.10   Median :156.0   Median :132.0  
##                     Mean   :60.92   Mean   :155.8   Mean   :132.2  
##                     3rd Qu.:64.20   3rd Qu.:162.0   3rd Qu.:139.0  
##                     Max.   :82.00   Max.   :176.0   Max.   :152.0  
##       dbp             hba1c             hcy             wt2       
##  Min.   : 71.00   Min.   : 2.400   Min.   : 8.80   Min.   :39.59  
##  1st Qu.: 82.00   1st Qu.: 5.800   1st Qu.:12.60   1st Qu.:52.09  
##  Median : 87.00   Median : 7.100   Median :14.20   Median :55.76  
##  Mean   : 86.54   Mean   : 7.048   Mean   :15.08   Mean   :58.19  
##  3rd Qu.: 91.00   3rd Qu.: 8.300   3rd Qu.:16.10   3rd Qu.:62.57  
##  Max.   :100.00   Max.   :11.000   Max.   :42.00   Max.   :81.54  
##       wt3             sbp2            sbp3            dbp2       
##  Min.   :39.43   Min.   :113.0   Min.   :111.0   Min.   : 62.00  
##  1st Qu.:51.25   1st Qu.:125.0   1st Qu.:125.0   1st Qu.: 77.00  
##  Median :55.11   Median :131.0   Median :130.0   Median : 82.00  
##  Mean   :57.61   Mean   :131.6   Mean   :130.7   Mean   : 82.31  
##  3rd Qu.:61.85   3rd Qu.:138.0   3rd Qu.:137.0   3rd Qu.: 87.00  
##  Max.   :81.07   Max.   :152.0   Max.   :153.0   Max.   :102.00  
##       dbp3           hpt            glucontrol
##  Min.   :67.00   Length:153         Good: 51  
##  1st Qu.:76.00   Class :character   Poor:102  
##  Median :81.00   Mode  :character             
##  Mean   :81.15                                
##  3rd Qu.:86.00  
##  Max.   :98.00  

#bmistatus (WHO classification)

healthstatcatbmi<- healthstatcat %>%
  mutate(height_m = ht / 100,bmi = wt / (height_m^2))

View(healthstatcatbmi)

healthstatcatbmi$bmistatus<- cut(healthstatcatbmi$bmi,
                           breaks=c(-Inf, 18.49999, 24.9999, 29.9999, Inf),
                           labels=c("underweight", "normal", "overweight", "o
bese"))
summary(healthstatcatbmi)

##        id            age             sex              exercise        
##  Min.   :  1   Min.   :21.00   Length:153         Length:153        
##  1st Qu.: 39   1st Qu.:36.00   Class :character   Class :character  
##  Median : 77   Median :42.00   Mode  :character   Mode  :character  
##  Mean   : 77   Mean   :42.16                                        
##  3rd Qu.:115   3rd Qu.:47.00                                        
##  Max.   :153   Max.   :64.00                                        
##   smoking              wt              ht              sbp       
##  Length:153         Min.   :42.60   Min.   :140.0   Min.   :114.0  
##  Class :character   1st Qu.:55.40   1st Qu.:148.0   1st Qu.:126.0  
##  Mode  :character   Median :59.10   Median :156.0   Median :132.0  
##                     Mean   :60.92   Mean   :155.8   Mean   :132.2  
##                     3rd Qu.:64.20   3rd Qu.:162.0   3rd Qu.:139.0  
##                     Max.   :82.00   Max.   :176.0   Max.   :152.0  
##       dbp             hba1c             hcy             wt2       
##  Min.   : 71.00   Min.   : 2.400   Min.   : 8.80   Min.   :39.59  
##  1st Qu.: 82.00   1st Qu.: 5.800   1st Qu.:12.60   1st Qu.:52.09  
##  Median : 87.00   Median : 7.100   Median :14.20   Median :55.76  
##  Mean   : 86.54   Mean   : 7.048   Mean   :15.08   Mean   :58.19  
##  3rd Qu.: 91.00   3rd Qu.: 8.300   3rd Qu.:16.10   3rd Qu.:62.57  
##  Max.   :100.00   Max.   :11.000   Max.   :42.00   Max.   :81.54  
##       wt3             sbp2            sbp3            dbp2       
##  Min.   :39.43   Min.   :113.0   Min.   :111.0   Min.   : 62.00  
##  1st Qu.:51.25   1st Qu.:125.0   1st Qu.:125.0   1st Qu.: 77.00  
##  Median :55.11   Median :131.0   Median :130.0   Median : 82.00  
##  Mean   :57.61   Mean   :131.6   Mean   :130.7   Mean   : 82.31  
##  3rd Qu.:61.85   3rd Qu.:138.0   3rd Qu.:137.0   3rd Qu.: 87.00  
##  Max.   :81.07   Max.   :152.0   Max.   :153.0   Max.   :102.00  
##       dbp3            hpt            glucontrol    height_m          bmi       
##  Min.   :67.00   Length:153         Good: 51   Min.   :1.400   Min.   :15.
65  
##  1st Qu.:76.00   Class :character   Poor:102   1st Qu.:1.480   1st Qu.:22.
06  
##  Median :81.00   Mode  :character              Median :1.560   Median :24.
```

```
89
##  Mean   :81.15                                 Mean   :1.558   Mean   :25.
31
##  3rd Qu.:86.00                                 3rd Qu.:1.620   3rd Qu.:28.
22
##  Max.   :98.00                                 Max.   :1.760   Max.   :38.
88
##       bmistatus
##  underweight: 6
##  normal     :75
##  overweight :48
##  obese      :24
##
##
```

```
#Reporting your descriptive analysis

library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary St
atistics Tables.

##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
stargazer(healthstatcatbmi)
```

```
##
## % Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University.
E-mail: hlavac at fas.harvard.edu
## % Date and time: Wed, Dec 16, 2020 - 11:34:00 AM
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lccccccc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
## Statistic & \multicolumn{1}{c}{N} & \multicolumn{1}{c}{Mean} & \multicolum
n{1}{c}{St. Dev.} & \multicolumn{1}{c}{Min} & \multicolumn{1}{c}{Pctl(25)} &
\multicolumn{1}{c}{Pctl(75)} & \multicolumn{1}{c}{Max} \\
## \hline \\[-1.8ex]
## id & 153 & 77.000 & 44.311 & 1 & 39 & 115 & 153 \\
## age & 153 & 42.163 & 8.932 & 21 & 36 & 47 & 64 \\
## wt & 153 & 60.920 & 8.270 & 42.600 & 55.400 & 64.200 & 82.000 \\
## ht & 153 & 155.797 & 8.885 & 140 & 148 & 162 & 176 \\
## sbp & 153 & 132.248 & 7.957 & 114 & 126 & 139 & 152 \\
## dbp & 153 & 86.536 & 6.268 & 71 & 82 & 91 & 100 \\
## hba1c & 153 & 7.048 & 1.785 & 2.400 & 5.800 & 8.300 & 11.000 \\
## hcy & 153 & 15.078 & 4.699 & 8.800 & 12.600 & 16.100 & 42.000 \\
## wt2 & 153 & 58.191 & 9.257 & 39.590 & 52.090 & 62.570 & 81.540 \\
## wt3 & 153 & 57.611 & 9.351 & 39.430 & 51.250 & 61.850 & 81.070 \\
## sbp2 & 153 & 131.575 & 8.103 & 113 & 125 & 138 & 152 \\
## sbp3 & 153 & 130.699 & 8.349 & 111 & 125 & 137 & 153 \\
## dbp2 & 153 & 82.314 & 7.514 & 62 & 77 & 87 & 102 \\
## dbp3 & 153 & 81.150 & 6.297 & 67 & 76 & 86 & 98 \\
## height\_m & 153 & 1.558 & 0.089 & 1.400 & 1.480 & 1.620 & 1.760 \\
## bmi & 153 & 25.315 & 4.285 & 15.647 & 22.056 & 28.217 & 38.877 \\
## \hline \\[-1.8ex]
## \end{tabular}
## \end{table}
```

```
stargazer(healthstatcatbmi, type = "html",
          title="Descriptive statistics",
          digits=1, out="table1.doc")
```

```
##
## <table style="text-align:center"><caption><strong>Descriptive statistics</
strong></caption>
## <tr><td colspan="8" style="border-bottom: 1px solid black"></td></tr><tr><
td style="text-align:left">Statistic</td><td>N</td><td>Mean</td><td>St. Dev.<
/td><td>Min</td><td>Pctl(25)</td><td>Pctl(75)</td><td>Max</td></tr>
## <tr><td colspan="8" style="border-bottom: 1px solid black"></td></tr><tr><
td style="text-align:left">id</td><td>153</td><td>77.0</td><td>44.3</td><td>1
</td><td>39</td><td>115</td><td>153</td></tr>
## <tr><td style="text-align:left">age</td><td>153</td><td>42.2</td><td>8.9</
td><td>21</td><td>36</td><td>47</td><td>64</td></tr>
## <tr><td style="text-align:left">wt</td><td>153</td><td>60.9</td><td>8.3</t
d><td>42.6</td><td>55.4</td><td>64.2</td><td>82.0</td></tr>
## <tr><td style="text-align:left">ht</td><td>153</td><td>155.8</td><td>8.9</
td><td>140</td><td>148</td><td>162</td><td>176</td></tr>
## <tr><td style="text-align:left">sbp</td><td>153</td><td>132.2</td><td>8.0<
/td><td>114</td><td>126</td><td>139</td><td>152</td></tr>
## <tr><td style="text-align:left">dbp</td><td>153</td><td>86.5</td><td>6.3</
td><td>71</td><td>82</td><td>91</td><td>100</td></tr>
## <tr><td style="text-align:left">hba1c</td><td>153</td><td>7.0</td><td>1.8<
/td><td>2.4</td><td>5.8</td><td>8.3</td><td>11.0</td></tr>
## <tr><td style="text-align:left">hcy</td><td>153</td><td>15.1</td><td>4.7</
td><td>8.8</td><td>12.6</td><td>16.1</td><td>42.0</td></tr>
## <tr><td style="text-align:left">wt2</td><td>153</td><td>58.2</td><td>9.3</
td><td>39.6</td><td>52.1</td><td>62.6</td><td>81.5</td></tr>
## <tr><td style="text-align:left">wt3</td><td>153</td><td>57.6</td><td>9.4</
td><td>39.4</td><td>51.2</td><td>61.9</td><td>81.1</td></tr>
## <tr><td style="text-align:left">sbp2</td><td>153</td><td>131.6</td><td>8.1
</td><td>113</td><td>125</td><td>138</td><td>152</td></tr>
## <tr><td style="text-align:left">sbp3</td><td>153</td><td>130.7</td><td>8.3
</td><td>111</td><td>125</td><td>137</td><td>153</td></tr>
## <tr><td style="text-align:left">dbp2</td><td>153</td><td>82.3</td><td>7.5<
/td><td>62</td><td>77</td><td>87</td><td>102</td></tr>
## <tr><td style="text-align:left">dbp3</td><td>153</td><td>81.2</td><td>6.3<
```

```
/td><td>67</td><td>76</td><td>86</td><td>98</td></tr>
## <tr><td style="text-align:left">height_m</td><td>153</td><td>1.6</td><td>0
.1</td><td>1.4</td><td>1.5</td><td>1.6</td><td>1.8</td></tr>
## <tr><td style="text-align:left">bmi</td><td>153</td><td>25.3</td><td>4.3</
td><td>15.6</td><td>22.1</td><td>28.2</td><td>38.9</td></tr>
## <tr><td colspan="8" style="border-bottom: 1px solid black"></td></tr></tab
le>

#for categorical data, just edit and add in the current table
#limitation in the r package stargazer
```